



Uncertainty in Gridded CO₂ Emissions Estimates

By: **Gregg Marland**, Eric Marland, Susannah Hogue, Robert J. Andres, and Dawn Woodard

Abstract

We are interested in the spatial distribution of fossil-fuel-related emissions of CO₂ for both geochemical and geopolitical reasons, but it is important to understand the uncertainty that exists in spatially explicit emissions estimates. Working from one of the widely used gridded data sets of CO₂ emissions, we examine the elements of uncertainty, focusing on gridded data for the United States at the scale of 1° latitude by 1° longitude. Uncertainty is introduced in the magnitude of total United States emissions, the magnitude and location of large point sources, the magnitude and distribution of non-point sources, and from the use of proxy data to characterize emissions. For the United States, we develop estimates of the contribution of each component of uncertainty. At 1° resolution, in most grid cells, the largest contribution to uncertainty comes from how well the distribution of the proxy (in this case population density) represents the distribution of emissions. In other grid cells, the magnitude and location of large point sources make the major contribution to uncertainty. Uncertainty in population density can be important where a large gradient in population density occurs near a grid cell boundary. Uncertainty is strongly scale-dependent with uncertainty increasing as grid size decreases. Uncertainty for our data set with 1 grid cells for the United States is typically on the order of $\pm 150\%$, but this is perhaps not excessive in a data set where emissions per grid cell vary over 8 orders of magnitude.

1. Introduction

There is a wide range of interest (both geochemical and geopolitical) in geographically explicit inventories of the sources and sinks of the greenhouse gas CO₂. It is a challenge to estimate sources and sinks in a spatially explicit context, and to best characterize the location and magnitude of emissions and sinks, we would like to estimate also the associated uncertainty. Emissions have rarely been measured directly (although there is an increasing amount of recent data on large point sources), and hence current gridded inventories of emissions from fossil-fuel use and industrial processes rely heavily on related, proxy, and re-purposed data. In the following analyses, we refine and combine the components of uncertainty and discuss them in the context of the widely used Carbon Dioxide Information Analysis Center (CDIAC) gridded inventory for fossil-fuel related emissions, focusing on the data from the United States [Andres *et al.*, 2013; see also Andres *et al.*, 2014]. We call on various statistics on large point sources and a variety of data relevant to the distributed and areal sources.

Few studies have explored the uncertainty of global-scale, grid-level emissions data sets. Rayner *et al.* [2010] noted that “none of the pointwise fossil emission products available today include” estimates of uncertainty. They then proceeded to try to develop uncertainty estimates for their data product. Although their data set is considerably different than the one discussed here in that they did not treat large point sources separately, they estimated that for their data set “uncertainties can be as high as 50% at the pixel level.” Relevant to our data set, Rayner *et al.* “described the complex structure of uncertainty that arises from combining pointwise and area-integrated constraints” and pointed out, importantly, that uncertainties for nearby pixels are not independent. Because, for example, the uncertainty for any given grid space includes consideration that a large point source might be only slightly displaced, they noted that “the ensemble statistics of many points are more reliable than those for an individual point,” a conclusion that we explore below by briefly examining the effect of changing grid size. Rayner *et al.* noted that “using the uncertainty of this pointwise map alone in an inversion is a serious error since it assumes independence of errors.”

Andres et al. [2012] also addressed the uncertainty of gridded emissions data and noted that using proxy data such as population density to distribute emissions begins to break down at finer scales and at low levels (e.g., in thinly populated areas). They reported that the percent error for grid spaces tends to increase where absolute emissions are highest. *Andres et al.* [2012] also added the warning that uncertainties in nearby cells are not independent.

With respect to spatial uncertainty, *Nassar et al.* [2013] reported simply that “Quantifying uncertainties in these data products ... is a challenge.” *Oda and Maksyutov* [2011] presented a lengthy discussion of uncertainty that focused on the problems in dealing with the locational uncertainty of large point sources and the temporal variability of large point sources. We note that reliance on data from large point sources places limitations on the development of long data time series or of disaggregating data to estimate emissions at time scales other than that for which sampling is reported.

2. Significance of Large Point Sources

Large point sources make up a large percentage of anthropogenic carbon dioxide emissions for the United States and for other industrialized countries [*Singer et al.*, 2014]. In 2010, one third of United States emissions were reported from only 311 sites of large point sources [*U.S. Environmental Protection Agency*, 2014]. We started by looking at data on large point sources from both the eGRID [*U.S. EPA*, 2014] and CarMA [*CarMA*, 2013] data sets. The eGRID data set from the *U.S. EPA* [2014] maintains information on emissions values and locations for all electric power plants in the United States, and *CarMA* [2013] uses this as input in developing a similar data set for the globe. As soon as the first few latitude and longitude data points from these data sets were typed into Google Earth many of these point sources were not observed at their reported locations. It was apparent that we would have to deal with both magnitude and locational uncertainty. Total uncertainty in emissions from any geographic grid space thus has to reflect uncertainty in small or areal sources and in both the magnitude and location of large point sources.

3. Uncertainty in Gridded CO₂ Emissions Inventories

Woodard et al. [2015] have developed one key component of what we need to quantify the spatially explicit uncertainty in gridded inventories of CO₂ emissions—an approach for dealing with the uncertainty in the locations of large point sources. This article proceeds to develop an approach for estimating and aggregating the impact of the other principal components of uncertainty.

While there are many gridded inventories that document ground level sources of anthropogenic emissions of CO₂ for the United States and the globe, the most commonly used are CDIAC, ODIAC, EDGAR, FFDAS, and Vulcan (see *Hutchins et al.* [2016] for a summary of each of these data sets and their differences). These inventories use a variety of different top-down and bottom-up (i.e., Vulcan) methods to geographically distribute emission on various sized geographic grids. Each of the top-down inventories uses some sort of proxy, such as population density or satellite-observed nightlights, to help distribute emissions totals from a large (national) scale down to the level of grids as small as 0.1° on a side. Some of the inventories use multiple proxies to take advantage of their differing characteristics.

However, using proxy data can result in the misallocation of emissions values both spatially and temporally. *Hutchins et al.* [2016] show evidence, through a simple comparison of the various data products for the United States, that the use of different proxies results in quite different allocations of emissions. *Hutchins et al.* show further that these differences increase as grid size is decreased. To begin to address this issue, we have taken the first steps toward calculating the total uncertainty for a CDIAC-like inventory for the United States at the 1° by 1° grid scale. Data here are estimates of annual emissions for the year 2009. Uncertainty values are for individual grid cells and do not consider the correlation of errors in adjacent grid spaces.

4. Uncertainty for a Modified, Gridded CDIAC Data Set

In the gridded CDIAC inventory [*Andres et al.*, 2013], data on population density are used as a proxy for the spatial distribution of all emissions within a country. For this analysis, we have removed emissions from electric power plants from the country total prior to using the population proxy to distribute the remaining national emissions. The power plants, with magnitudes and locations from EPA's eGRID [*U.S. EPA*, 2014] data

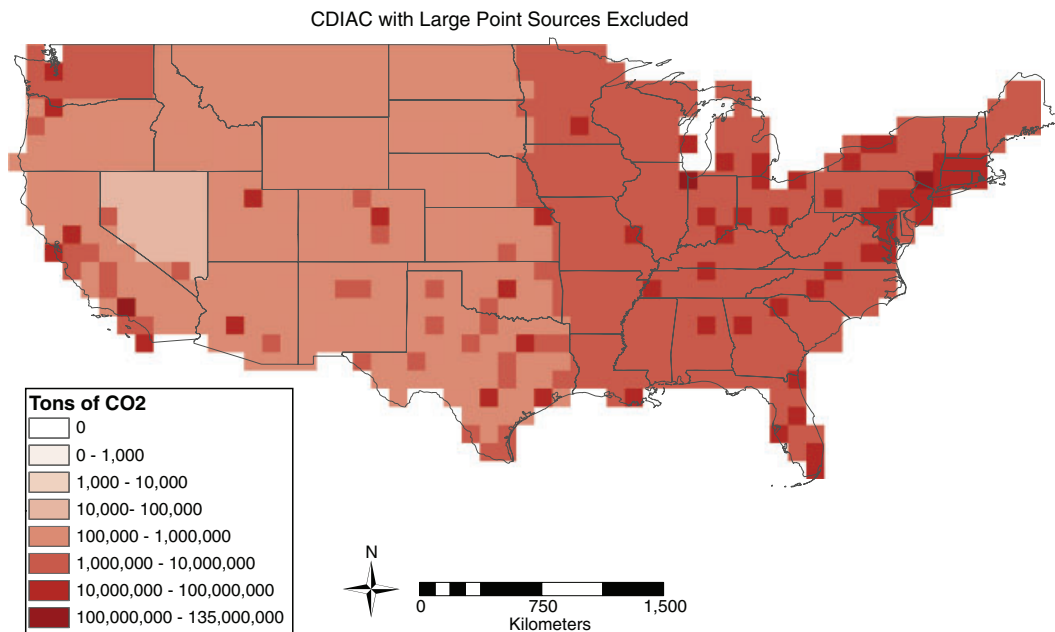


Figure 1. The Carbon Dioxide Information Analysis Center data set of CO₂ emissions for the United States in 2009 with total emissions from large point sources removed from the national total. Population density was then used as a proxy to distribute the remaining emissions on a 1° grid, as shown here. Emissions are shown in metric tons of CO₂.

set, were then added back to give total emissions in each grid cell. Figure 1 shows the inventory of emissions from the United States with emissions from power plants not included.

The emissions inventory discussed here is thus a derivative of the CDIAC data set and is comprised of two components, power plant emissions from the eGRID data set and all remaining national emissions distributed using population density as a proxy. These remaining emissions do contain some additional, large industrial sources of CO₂, but reporting to the EPA GHG Reporting Program [U.S. Environmental Protection Agency, 2015a] shows that in 2010, 73% of emissions from sources greater than 25,000 t of CO₂ equivalent were from power plants. Comparable data on industrial sources are not available outside of the United States, and for the purposes of this study, we assume these industrial sources to be part of the areal sources of emissions (hereafter “non-point sources”).

We thus consider the six major components of uncertainty that need to be combined for an initial estimate of uncertainty for the cells in the modified CDIAC database:

- Uncertainty in total national emissions.
- Magnitude uncertainty for large point sources.
- Spatial uncertainty for large point sources.
- Magnitude uncertainty of the population proxy.
- Spatial uncertainty for the population proxy.
- Uncertainty in using population density as a proxy for emissions.

For this analysis, all calculations are based on one-sigma uncertainty. In combining the component uncertainties, we assume that all of the components above, except the uncertainty in the national total, are independent of each other and are represented in units of metric tons of carbon dioxide. As they are assumed to be independent, they are then combined in Euclidean fashion (the square root of the sum of the squares) on a grid cell by grid cell basis. The national total uncertainty is included as a multiplier factor because all other values depend implicitly on it. In this analysis of uncertainty, we use data for emissions in 2009 as published by CDIAC in 2013. Since we are only considering 1 year of data, we do not consider temporal uncertainty. Temporal uncertainty would need to be considered in developing a time series of emissions.

One technical note on the uncertainty calculations is needed before we begin. The uncertainty values calculated below are frequently larger than 100% which makes a symmetric uncertainty unrealistic. All of the

values below should be thought of on the lower side as the smaller of the calculated uncertainty and the magnitude of the value in question.

4.1. Uncertainty in Total National Emissions

Uncertainty in total, national U.S. emissions is estimated at 2.5% (one sigma). This value is based on comparisons with other inventories, U.S. EPA analyses, and literature research of the components, which combine to calculate total national U.S. emissions [see also *Andres et al.*, 2014]. According to the *U.S. Environmental Protection Agency* [2015b], the 95% confidence interval for total U.S. CO₂ emissions from fossil-fuel combustion is -2% to +5%. Instead of using this asymmetric value, we take the symmetric value of $\pm 5\%$ and because this is two standard deviations about the mean and our computations are all based on one standard deviation, the estimated national error used in our computations is $\pm 2.5\%$. We assume the national U.S. emissions uncertainty estimate is applicable to the 48 contiguous states considered here.

4.2. Magnitude and Spatial Uncertainty of Emissions From Large Point Sources

A random sample of 500 large point sources from the U.S. EPA eGRID data set was taken in order to find the exact locations of the power plant discharges. We used Google Earth satellite imagery to identify the point sources and to verify each latitude and longitude. With spatial information from Google Earth, the distance between the actual location of CO₂ emissions and the reported location was computed for each point source in the sample. The maximum distance from the reported location to the observed location of a point source was approximately 106 km. The mean distance from the reported location for all of the sample point sources (excluding zero) was 1.97 km. The mean distance from the reported locations for all the point sources in the sample was 0.84 km. The latter value was then used as the mean spatial uncertainty [Woodard *et al.*, 2015]. Given the inaccuracy associated with the 500-item sample of point sources, we took a closer look at the largest emitters. The spatial uncertainty for the top 81 emitters was larger than that for the random sample of points. It was found that 60% were farther than 1 km from the reported location. The mean difference in location was 7.94 km and the maximum spatial difference was about 122 km. This suggests that there is no better information on the larger facilities, and we assume that locational uncertainty is independent of facility size.

The information gathered from the 500-item random sample suggested that the differences between discharge locations and eGRID reported that the locations could be variously attributed to: (1) differences between the plant site as opposed to the exact location of the CO₂ discharge stack, (2) using a default location in the EPA database, such as the centroid of a county, when the initial report to EPA did not include plant coordinates, (3) reporting the location of a company office or mailing address instead of the power plant site, (4) dealing with the existence of multiple stacks on the same site, and (5) typographical errors.

The locations of power plants with respect to a given grid space are not part of a continuous distribution, and therefore, most traditional statistical methods do not work well in dealing with the uncertainty in their emissions. The discrete, or binary, nature of the locations (a plant either is in a given grid space or it is not) spurred the creation of a new method for dealing with the likely locations and the uncertainty in the emissions from power plants. A complete treatment is given in Woodard *et al.* [2015], but a brief overview is included here.

A Monte Carlo simulation was used to distribute the emissions from each reported point source according to an assumed form for the distribution of its likely location given its reported location. The random-sample data provided the distribution of distances from the reported location to the actual location. Assuming no direction bias, a radially symmetric bivariate normal distribution was used to simulate the location of the point source around the reported location. The mean distance from the reported location was matched to the mean distance found in the random-sample data. The points from the Monte Carlo distribution were overlaid and allocated to grid cells at our specified resolution. This provides what is effectively an expected value of emissions in the grid cells in a region around the reported location. The closer that the point source is reported to the center of the grid cell the higher the probability that the point source will actually be found in the reported cell.

In addition, the random-sample data suggested that the county of origin of the reported point sources was virtually always correct. Using this information, all points in the simulation falling outside of the reported county of a point source were not counted. Relatedly, if the coordinates of a reported point source were not

consistent with the reported county, the point source was placed at the centroid of the county—again recognizing that the large point sources were essentially always identified in the correct county. This approach is consistent with the database’s documented approach for placing point sources with missing locational coordinates.

The technical notes on the reported data set suggested that point sources with missing locational data were allocated by default to the centroid of the county in which they were reported. With the sample data confirming this feature, that point sources reported in the centroid of a county had much larger uncertainty (might actually lie anywhere in the county), the distribution used in the Monte Carlo simulation for points located at the centroid of a county was given a mean uncertainty matching the mean distance between the centroid of the county and its perimeter. The result is that point sources reported in the centroid of a county have much greater uncertainty and a larger spread of emissions than point sources reported at other locations within a county.

In summary, the simulation computes for each point source the mean simulated emissions, analogous to the expected value, for each grid cell in the neighborhood of the reported grid cell location. The result provides the simulated mean value of emissions for each grid cell. The resulting grid of simulated means effectively distributes the reported CO₂ emissions from a point source to surrounding cells based on the fraction of the total number of simulation executions that fell in each cell (Figure 2).

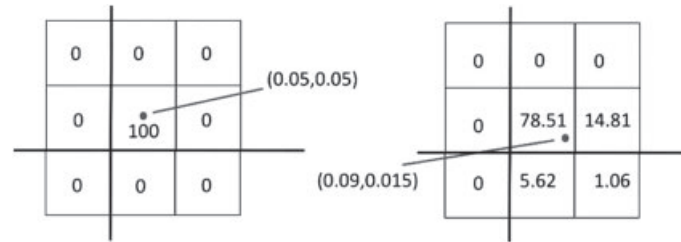


Figure 2. This is a simplified illustration of how emissions totaling 100 t of CO₂ result in a grid of simulated means. The left panel shows how simulated emissions would be distributed if the point source is reported in the center of a grid cell and the locational uncertainty (0.84 km) is small enough that there is very low probability that the point source is actually located outside of the reported cell. The right panel shows how the simulated emissions would be distributed if the point source is off center and the uncertainty is large enough (still 0.84 km) that there is meaningful likelihood that the actual point lies in an adjacent cell. The total emissions attributed to the source remain the same, but the distribution among the reported cell and surrounding cells depends on the location within the central cell and the shape of the bivariate normal distribution of uncertainty. Numbers in parentheses show the coordinates of the power plant within cells that are 0.1° × 0.1° of latitude and longitude and are at a latitude characteristic of the central United States. If the cells were 1.0° by 1.0° and the locational uncertainty still 0.84 km, both power plants would be 100% retained in the reported cell.

To compute the uncertainty in the reported emissions from a point source, a comparison was made between the reported emissions and the distributed emissions from the simulation. Because of the binary nature of the emissions at a large point source, classic spatial methods based on smooth distributions did not apply or provide useful results. Instead, we derived a new statistic that parallels the approach of a standard deviation, but provides a more useful representation of the uncertainty.

The new statistic, Point Source Uncertainty Measure (PSUM), is defined as the root mean square difference between the simulated

means and the reported emissions value [see Woodard *et al.*, 2015]. That is, PSUM calculates a measure of the difference between the reported emissions from a grid cell and the distributed emissions from the simulation in that grid cell.

PSUM is formally defined as the square root of the sum of the squares of the weighted differences in the reported and simulated emissions values.

$$PSUM = \sqrt{\sum_k p_k (s_k - x_{ij})^2}$$

The x_{ij} are the means calculated in the simulation and the s_k are the reported values (either 0 or M , with M being the emissions value) for each point source. The frequency of occurrence, p_k , completes the calculation.

This is essentially an average where each entry is weighted according to its frequency of occurrence. Each outcome in the simulation is either M or 0 because we are taking the difference between the simulation value and the reported value at each grid cell. There are only two frequencies that need to be computed,

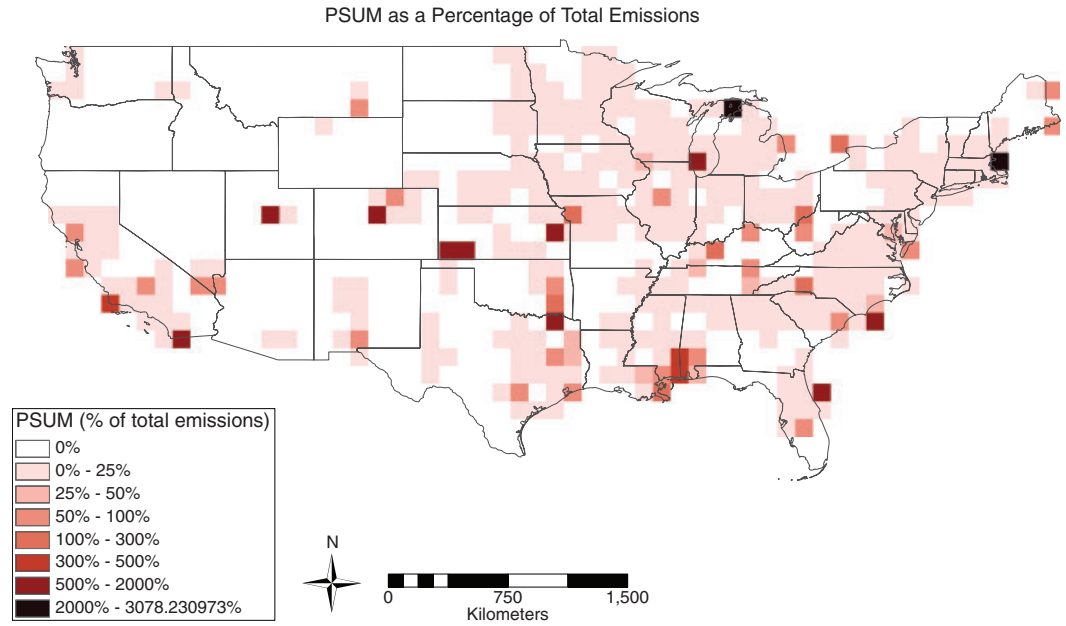


Figure 3. Point Source Uncertainty Measure (PSUM) output is shown as a percentage of total emissions in each grid space from large point sources at 1° resolution. This image includes the refinements made to PSUM regarding point sources that are reported in the centroid of their county (see text). PSUM is not strictly a measure of one standard deviation of uncertainty but it has similar characteristics, it was designed to treat the binary circumstance that a large point source either is or is not in a given grid space, and it is treated as such [see Woodard *et al.*, 2015].

the number of times that the output is M and the number of times it is 0. Thus, $p_k = \left\{ \frac{x_{ij}}{M}, 1 - \frac{x_{ij}}{M} \right\}$. The PSUM is then computed separately for each grid cell when the reported emissions value is M and when the reported emissions value is 0. PSUM can be treated similarly to a standard deviation and has the same units. Our measures of spatial uncertainty are thus the derived PSUM values as described in Woodard *et al.* [2015] (see Figure 3).

The magnitude uncertainty for emissions from large point sources is taken to be a constant $\pm 10.62\%$ (two sigma, one sigma = 5.31%) [Quick, 2014; see also Schindler, 2015]. This number was derived by comparing data collected on smokestack emissions of U.S. electric power plants with emissions calculated from fuel deliveries at the same plants. The mean value of the difference of the two data sets was 0.7% with the difference varying by 10.62% (two sigma) for individual plants.

4.3. Magnitude and Spatial Uncertainty in Population Density

There are multiple data sets that one might select to display the distribution of population density across the United States and ultimately the globe. LandScan is a global data set from Oak Ridge National Laboratory [2015] that integrates over time to estimate the average locations where people actually are rather than where their home location is. LandScan was first produced in 1998 and data sets for 2000–2012 are now available. CDIAC has contemplated use of LandScan population data but has not yet made the conversion. We use LandScan here to illustrate the uncertainty that could be achieved for the post-2000 time period. The CDIAC gridded CO₂ emissions data set relies on a population distribution data base from the Goddard Institute for Space Studies [see Andres *et al.*, 1996].

Magnitude uncertainty in LandScan for the United States was assumed to be comparable to the estimates of uncertainty derived by the U.S. Census Bureau at the same spatial scale [U.S. Census Bureau, 2011]. LandScan does not currently have any published estimates for uncertainty, but we assumed that it is very low in the United States. As in all of the data sets used here, the uncertainty will vary by country or region in a global analysis. The estimate provided by the U.S. Census Bureau is 0.01%, so this is the value used in our computations.

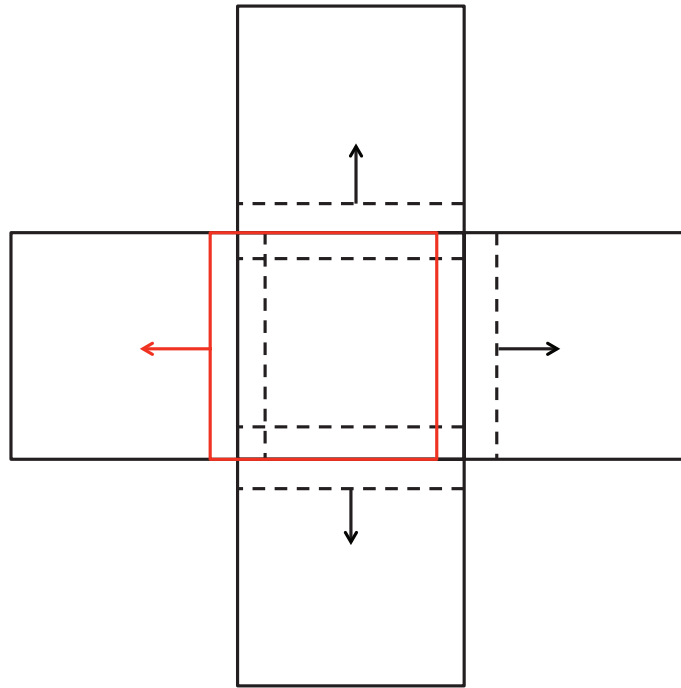


Figure 4. This diagram illustrates the shifting that occurs in order to compute the spatial uncertainty for a proxy data set like population density. The size of the shift is determined by a selected distance or by a selected fraction of the size of the grid cell. The size of the shift can be adjusted to the reliability of the data in use. This diagram depicts the shifting of a single grid cell in each of the four directions. The algorithm shifted all of the grid cells in each of the four directions.

Spatial uncertainty in LandScan was estimated by looking at the changes incurred as a result of small shifts in the cell boundaries. We started by taking the LandScan data set and distributing CO₂ emissions proportional to the population density values associated with each grid cell. We then shifted the grid cells by 0.1° (about 11 km in the continental United States) in each direction (N, S, E, and W) so that, for 1° grid cells, each grid cell contained successively one tenth of each of the four surrounding cells (see Figure 4). This effectively creates a weighted sum in which the central cell emissions value is weighted by 90% and the cell that is shifted toward the center is weighted by the remaining 10%. A weighted sum was computed for each of the four shifts that occurred. The standard deviation for the resulting

four weighted sums was then computed and stored as the uncertainty value within the central cell of a blank data set whose cell placement directly corresponds to the cells in the original data set (Figure 5). This is a preliminary uncertainty estimate that can and should be improved upon in later work.

4.4. Population Density as a Proxy for Emissions

In order to characterize the uncertainty associated with using population density as a proxy for CO₂ emissions, we must identify what information is known and can be used. We can find the mean number of grid cells per state, we know the population of each state, and we have information on state emissions from non-point sources, i.e., emissions totals for each state that excludes power plants [U.S. Census Bureau, 2011; U.S. Department of Energy/Energy Information Administration, 2013; U.S. EPA, 2014]. The basic premise of using population density as a proxy for CO₂ emissions is that per capita emissions are constant. Our basic assumption is that the variability in per capita emissions across states gives us an initial measure of the variability within states. From state-level data, we can calculate the mean level of per capita emissions for each state and then calculate the standard deviation in per capita emissions from non-point sources at the state level. We take this standard deviation as a measure of variability in the relationship between population density and emissions density at the state level. We assume then that the variability among states provides a basis to extrapolate the variability at the grid level within states. The state data provide enough information that we can back calculate to provide a rough estimate of the standard deviation in emissions by grid cell attributed to the population proxy. We use this standard deviation as a measure of the variability, i.e., the uncertainty estimate, when using population density as a proxy for emissions at the grid cell level.

Consider a large number of grid cells where we have computed the mean and variation in per capita emissions. We can group these cells into aggregates containing different numbers of cells and investigate the relationship between the size of the aggregate and the mean and variation in per capita emissions among those aggregates as compared to the mean and variation in the original ensemble of grid cells. As the aggregates, and hence, the number of grid cells in each aggregate, increase in size, the mean of the per capita emissions over those areas remains constant. The variation does not. This is a consequence of the Central

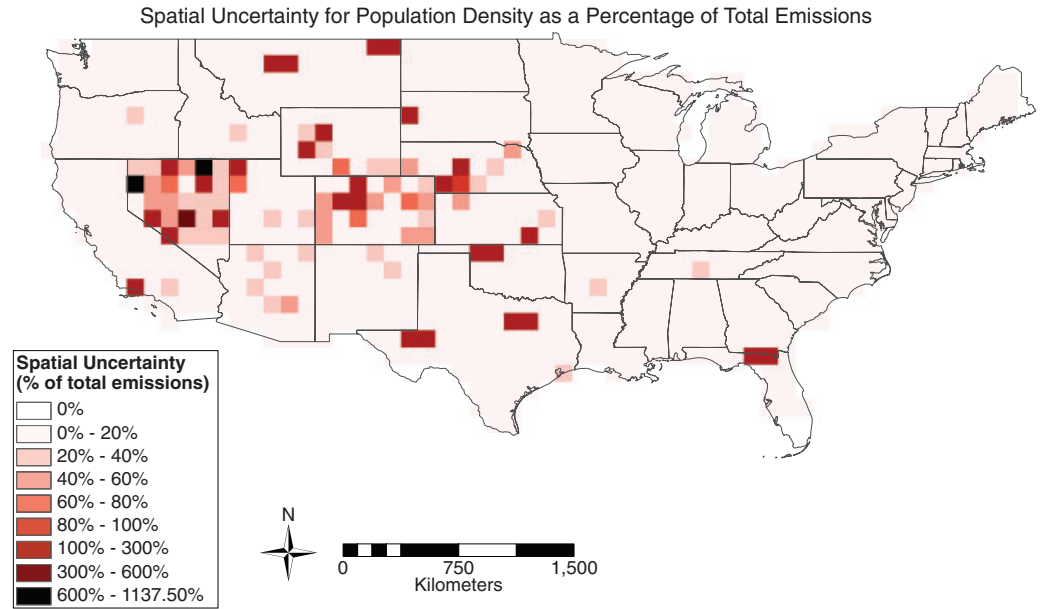


Figure 5. Spatial uncertainty associated with the spatial distribution of CO₂ emissions from non-point sources, based on the LandScan data set on population density, as a percentage of total emissions, shown at 1° resolution. This image was made by shifting each grid cell by 0.1° in each direction as shown in Figure 4. Each central cell was then weighted by 90% while the surrounding cells were weighted by 10%. Large values occur when cell borders approximately correspond with sharp changes in population density.

Limit Theorem from basic statistics. The Central Limit Theorem requires that the per capita emissions in the respective grid cells be independent from each other and at a fine scale we do not believe this to be true. However, at the scale of 1° grids (over 100 km on a side) and having removed power plant emissions, we believe that this is a reasonable first order assumption.

Under this assumption, the relationship between the number of grid cells in each aggregate and the standard deviation is such that the standard deviation is inversely proportional to the square root of the number of grid cells in each aggregate. Using this relationship, we use the variation in the per capita emissions at the state level to estimate the variation, or standard deviation, in per capita emissions at the grid level. The following formula allows us to make this estimate of the standard deviation at the grid cell level.

$$\sigma_{\text{state}} = \frac{\sigma_{\text{grid}}}{\sqrt{n}}$$

Here, σ_{state} is the standard deviation in emissions by state and σ_{grid} is the standard deviation in emissions by grid cell. This approach serves to compute a \pm value on areal emissions per cell. The application of this approach is shown in Figure 6. Per capita emissions for the United States as a whole were 8.4 t CO₂ per year for 2010 (using data from DOE/EIA for total emissions, the EPA for power plant emissions, and the Census Bureau for population). The per capita emissions for non-point sources by state vary from 4.5 to 23 t CO₂ per year (excluding Hawaii and Alaska but including Washington D.C.). The standard deviation around the 8.4 value is 3.25. The average number of grid cells per state is about 17.5. The relationship between the standard deviation at the grid level and at the average of 17.5 grid aggregates will be related by a factor of $\sqrt{17.5}$. This puts the standard deviation of per capita emissions at the grid level to be about 13.60 t CO₂ per person per year. Our estimate of the uncertainty in each grid space associated with using population density as a proxy for non-point-source CO₂ emissions is shown in Figure 6.

There are refinements that could allow for a more accurate result. This treatment essentially assumes that all states have the same number of grid spaces. Large states, for example, would create outliers in the data used to find the mean number of grid cells for each state. Although conceptually simple, but challenging in execution, if we could separate states into groups by relative size and treat these groups separately, we could get a more accurate idea of the standard deviation by grid cell for states of similar size. Also, being able to include large industrial sources along with power plants in the category of large point sources would

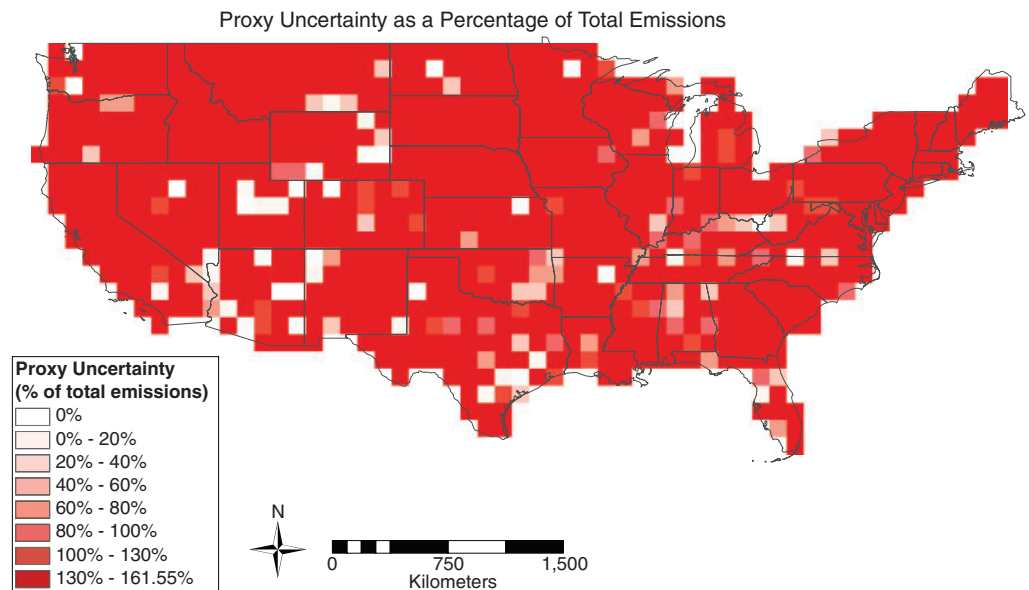


Figure 6. The uncertainty associated with using population density as a proxy for emissions is shown as a percentage of total emissions at 1° resolution. It was computed by scaling the variation in per capita emissions by state to the level of a single grid cell (see text).

allow for a more accurate computation of the standard deviation of the non-point-source emissions at the state level.

5. Combined Uncertainty for a Gridded Fossil-Fuel-Related CO₂ Emissions Data Set

The Euclidean sum of the five distributed aspects of uncertainty (the national total uncertainty is then a multiplier for the combined sum) produces a composite uncertainty, by grid space, for a hypothetical, modified CDIAC data set. Figure 7 shows the results of this summation for the United States for the year 2009 and Figure 8 shows these results as a percent of total emissions in the given grid space. Recall that these numbers apply to a hypothetical data set—one in which (1) the data on large point sources have been substituted for an equal quantity of emissions that were previously distributed according to population density and (2) the data on population density have an uncertainty attributed to data from the U.S. Census Bureau at the same scale. In both cases, we expect that the best achievable uncertainty will be larger in many other countries where the data on population and large point sources have greater uncertainty.

The scale shown on Figure 8 is very high. This was expected because of how much uncertainty there is for the exact location and magnitude for CO₂ sources at this scale. Figure 8 shows that the uncertainty associated with the modified CDIAC data set is consistently around 150% of the emissions total for each grid space. There are a few outliers, such as in Nevada, as a result of the high spatial uncertainty that forms around cities with high population density but very low population density in surrounding grid cells. Recall that these estimates of uncertainty are for a modified CDIAC data set where we have now isolated large point sources before using population density to distribute the remaining emissions from non-point sources. We have also treated the population density data as though they had been derived from Landsat, thus avoiding the shifts in population density that have occurred since construction of the Goddard Institute for Space Studies data set for 1984. Recall too that this uncertainty is for individual grid cells of 1° scale and is very scale-dependent. There is strong correlation among grid spaces because of the spatial uncertainty about the exact placement of large point sources and because the national total is a defined constant.

6. Analysis and Reduction of the Uncertainty

Total uncertainty is the combination of all of the components, but we also can learn something of the role that each component takes in forming the whole. With an understanding of the relative magnitude of each

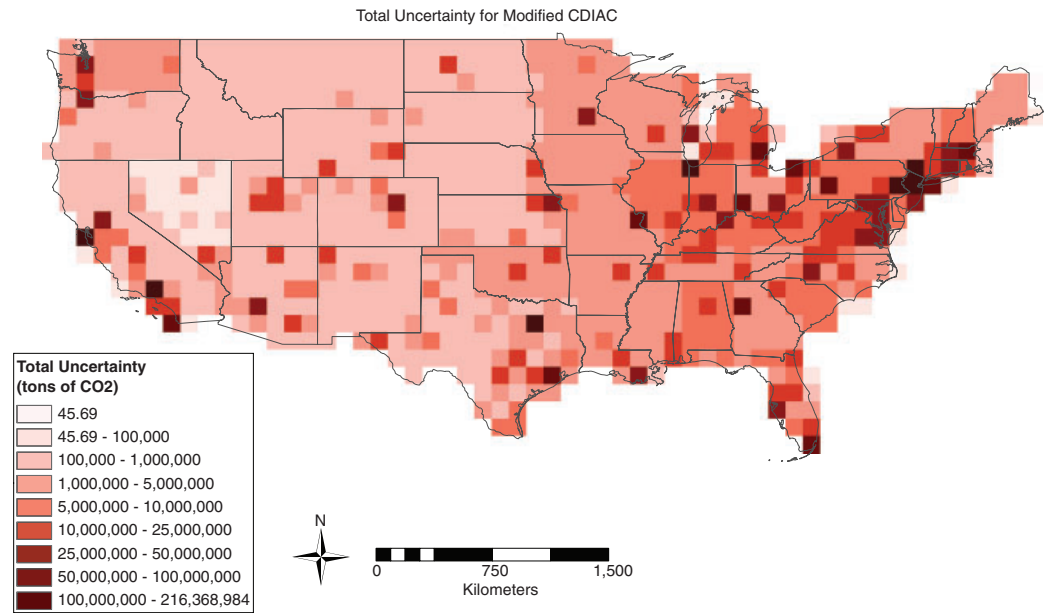


Figure 7. Total uncertainty, by grid space, for a hypothetical, modified Carbon Dioxide Information Analysis Center data set shown at 1° resolution, expressed as metric tons of CO₂. It is a simple Euclidian sum of five of the six components of uncertainty, with the uncertainty in the national total as a multiplier of this sum.

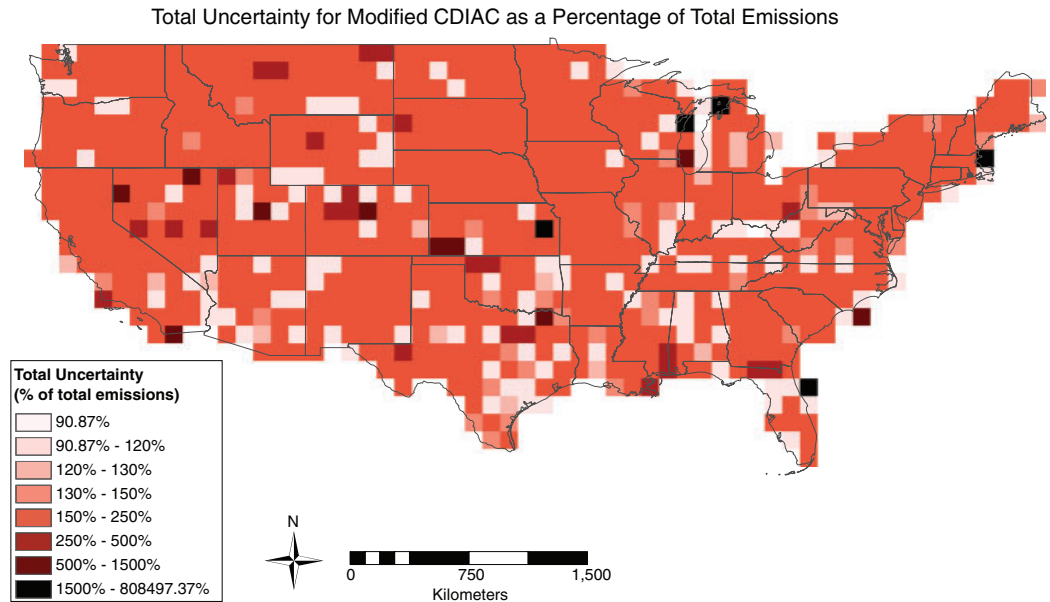


Figure 8. Uncertainty by grid space is shown as a percentage of total emissions at 1° resolution. Areas shown with very high uncertainty are often a result of cities with abrupt changes in population density. Excluding these few areas of very high uncertainty, we can see that the overall uncertainty in grid spaces is generally on the order of 1.5 times the total emissions. This is for a hypothetical, modified version of the Carbon Dioxide Information Analysis Center data set (see text).

of the pieces and the locational characteristics of where each component is large we can try to target specific efforts to best reduce the total uncertainty and to understand where those efforts might go wasted. A deeper understanding of the grid cell by grid cell uncertainty also provides a glimpse into the nature of the uncertainty in each component.

Our analysis suggests that for most 1° × 1° grid cells the uncertainty is between 160% and 170% of total emissions (66% of grid cells) (see Table 1). A small number of cells (5.7% of the total) have uncertainties

Table 1. This Is, by Grid Cell, a Breakdown of Each Component of Uncertainty With Its Summary Statistics					
	Minimum	1st Quartile	Median	3rd Quartile	Maximum
Magnitude uncertainty for large point sources	0.00	0.00	0.02	6.65	105.30
Spatial uncertainty for large point sources	0.00	0.00	0.00	0.01	784,900.00
Magnitude uncertainty for LandScan	0.00	0.01	0.01	0.01	0.01
Spatial uncertainty for LandScan	0.00	0.45	0.65	4.13	1142.00
Proxy uncertainty	0.0	151.3	161.5	161.5	161.5
Total ^a uncertainty for emissions data at grid cell level	90.4	160.9	165.6	165.7	804,600.0

Uncertainty in the national total is not included because it affects each of the grid cells equally. Note that the country borders create problems in that emissions may or may not occur even if a grid cell is designated as predominantly ocean, and some of the zero values lie along the eastern shoreline of the United States. This is one of the challenges of cropping a global data set to a single country for a targeted analysis. Units are the % of emissions in a grid cell.

^aThe total uncertainty includes the national level uncertainty of 2.5% whereas the other rows are only for the factor shown.

above 200% and an equally small number below 100% (6.1% of all grid cells). Because uncertainty is particularly high in some areas, we looked at characteristics that correspond with higher values of uncertainty. Table 1 shows summary values, as a percent of total emissions, in metric tons of CO₂, for the uncertainty associated with each component of the data set uncertainty. These statistics help identify which components have the greatest impact on the total uncertainty for the data set.

We can divide the uncertainty into the three basic pieces from which they arise: LandScan, the population proxy, and large point sources. Each of these plays an important role in the overall uncertainty and we look into each separately to see how they fit into the larger picture and where we might reduce their respective contributions. One message of Table 1 is that the magnitude uncertainty in the LandScan population density data is so small that it can almost be ignored. All of the other components have uncertainty values that reach up to 100% or more of the emissions levels. At the mean, proxy uncertainty is the largest contributor but there are a small number of grid spaces with very large spatial uncertainty from LandScan population density.

The spatial uncertainty from the LandScan population density data is particularly sensitive to abrupt changes in population density that occur near grid cell borders, where a small change in location can lead to large contributions to uncertainty. In most locations, the spatial uncertainty derived from LandScan is extremely low as shown both in Table 1 and in Figure 5, and Figure 5 shows that the number of grid cells where the values are high is quite small. There are only four grid spaces in Figure 5, all occurring in Nevada, where this uncertainty level reaches over 250%. At some cities with abrupt transition from high to low population density, such as Las Vegas and Reno, there is no such contribution to uncertainty because the transition from city to rural land does not occur near the edge of a grid cell. We note that many more of these sharp transitions would be expected at smaller grid sizes. More grid cells give more boundaries and more chances that these abrupt changes will happen close to a cell border.

Table 1 also indicates that proxy uncertainty has the highest median percentage of all the components. The pie chart in Figure 9 shows the percentage of the uncertainty attributable to proxy uncertainty by grid cell. In particular, over 83% of the grid cells have proxy uncertainty comprising over 80% of the total uncertainty. In fact, a full 79% of grid cells have 90% of their uncertainty coming from proxy uncertainty. The implication here is that in the majority of grid cells, at 1° grid size, reduction of uncertainty can only be carried out by addressing uncertainty in the proxy relationship. This means that we must obtain a better understanding of the relationship between the proxies we use and the emissions they are meant to represent.

One possible approach to reducing proxy error is to use sectoral data and/or bottom-up inventories such as Vulcan [Gurney *et al.*, 2009] to place more accurate bounds on the relationship between emissions and the various proxies being used in gridded inventories. Although these bottom-up inventories are more time and cost intensive, the value in understanding the uncertainties can be considerable. A challenge is that the relationship between bottom-up data and proxies likely varies among countries and regions, and in this

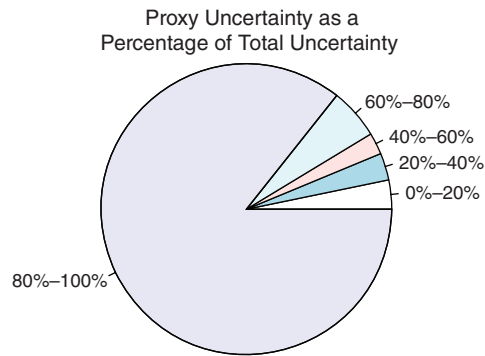


Figure 9. This pie chart shows that the majority of uncertainty values for the proxy uncertainty lie in the range of 80–100% of the total emissions values.

have such high emissions that even moderate uncertainties in the large point sources can dominate over the proxy uncertainties. The uncertainties from large point sources dominate the uncertainty in two ways, in two characteristic locations. The primary effect is the magnitude uncertainty of the large point sources in the reported locations of those sources. But there is also an important contribution in adjacent cells because of the spatial uncertainty and the possibility that the large point source might actually be in an adjacent cell. These two effects combined contribute to making up the majority of emissions uncertainty in over 11% of the grid cells in the U.S. Efforts to reduce these components of uncertainty rely on improving locational reporting of the large point sources and improving emissions measurements from these sources.

7. Discussion

7.1. Reducing Spatial Uncertainty of Emissions From Large Point Sources

Some point sources that are listed in the eGRID data set do not have a unique reported location. When reporting entities do not report a detailed location, the eGRID data system assumes that the point source is correctly located in the county that it is reported in and allocates it to the centroid of that county. This gives us the opportunity to better manage the locational uncertainty when we can identify that a power plant is located at the centroid of a county.

Collected data on point sources that were reported near county centroids help to clarify the uncertainty. A total of 137 point sources in eGRID were reported within 2 km of their county centroid. Each of these point sources was located using Google Earth satellite imagery and Google address searches to calculate the distance between their actual location and their reported location. Fifty-eight of these 137 point sources were reported in their actual locations and had a spatial distance of zero between the reported and actual locations. The remaining 79 point sources were not found in their reported locations. The 137 data points were then partitioned into groups according to their calculated spatial differences. Table 2 shows how the percentage of point sources reported in their exact location decreases as the computed distance from the centroid gets smaller. It also indicates that point

Table 2. Large Point Sources of CO₂ Emissions That Are Reported to be Close to the Centroid of a County Are Likely to be Shown With Default Locations Rather Than With Actual Locations

Distance From Reported Location to County Centroid (km)	Percent Reported in Actual Location (%)
0–0.5	9
0.5–1	22.5
1–1.5	60.5
1.5–2	69.7

By the time a reported location is as much as 1 km from the county centroid it is much more likely that the reported location corresponds to the actual location.

case sectoral data could be particularly useful. To balance minimizing expense with better characterizing uncertainty, the correlations gained by a bottom-up or detailed sectoral study might only need to be made periodically. But as gridded inventories move to smaller time intervals and smaller spatial scales, the effort needed to invest in a bottom-up or detailed sectoral approach may be great, and it may be possible to use periodic detailed inventories to help calibrate the proxy relationship.

Note that the locations where proxy uncertainty is dominant are not the locations where the largest emissions occur. Proxy uncertainty dominates in locations where there are no large point sources. The large point sources

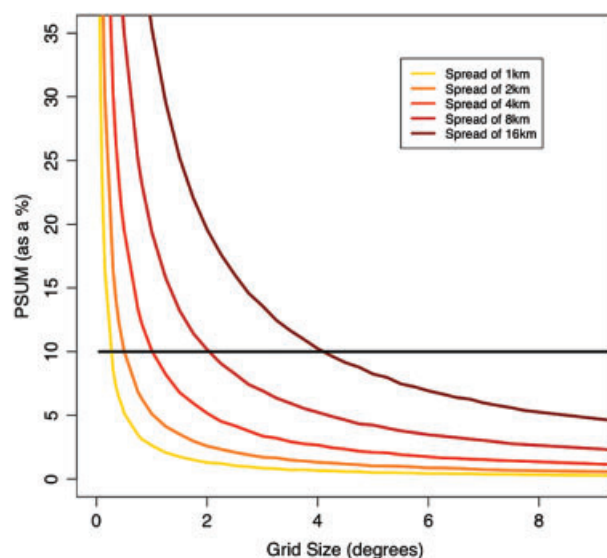


Figure 10. This simulation, based on the spatial uncertainty of large point sources of emissions, shows that uncertainty in gridded emissions values decreases as grid size increases. This provides a basis for determining a grid resolution that corresponds to the accuracy of the data being used. Differences in the "spread" are the mean difference between the reported and observed locations of large point sources.

sources within 1 km of the centroid of the county are likely to be reported to the incorrect, i.e., default, location.

The placement of point sources in county centroids occurs often enough that it is useful to note the decline in accuracy for point sources reported to be at the centroid of a county. Of all the data points found within 2 km of the centroid of their county, 42% were reported in their actual location. Thus, the annuli within the range of 0–1 km have a higher percentage of point sources being reported to a default location rather than to the actual location. This gives reason to treat point sources reported within a 1 km radius of the centroid of a county with different spatial error. The metric for spatial uncertainty is thus increased so that the range of possibility is the entire county instead of the standard 0.84 km mean spatial uncertainty that is used in the initial PSUM simulation.

The mean spatial uncertainty for eGRID, excluding the points that fall within 1 km of a county centroid, is about 0.74 km. Additional characterizations of power plants and other large point sources could be used to further narrow and refine uncertainties. We know, for example, that coal power plants need significant water sources for cooling. Intersecting maps of water sources with reported plant locations could allow additional reductions in uncertainty for coal-powered plants by recognizing that those distant from water sources are more likely to be mislocated. This kind of analysis would allow the uncertainty of remaining points to be reduced and might have important implications for countries with less certain data sets.

7.2. A Relationship Between Grid Scale and Uncertainty

An initial sampling of data on large point sources suggests that spatial uncertainty varies significantly among countries and regions. The implication of this is that either the grid-space uncertainty will vary greatly among regions or that the "resolution" of the resulting plots would need to be adjusted across regions to maintain the same level of uncertainty. Figure 10 shows how the spatial uncertainty measure for large point sources might lead to data reporting on different scales (grid sizes). The figure shows a horizontal line plotted at 10% uncertainty (PSUM). The relationship between PSUM and grid size is then shown for various levels of the mean value of the distance between reported and actual locations of large point sources. For high values of spatial separation, a larger grid size is required to restrain PSUM to the same level of uncertainty. Other components of the uncertainty discussion will be sensitive similarly to changes in scale. The issue of spatial resolution is important for emissions data products. It is not always clear what resolution means, but at some level, we can understand that reporting at very small scales does not mean that the resolution is increased usefully because uncertainty increases as resolution increases. Uncertainty must be incorporated into our idea of how resolution is employed. *Hutchins et al. [2016]* show that the different data sets on gridded CO₂ emissions estimates converge very rapidly as grid size is increased.

8. Conclusions

CO₂ emissions from fossil-fuel use play a major role in global climate change and for both geochemical and geopolitical reasons, it is important to understand the origins of CO₂ emissions. Understanding anthropogenic emissions provides a boundary condition for unraveling the dynamics of the global carbon cycle

and for predicting its behavior into the future. There remains considerable uncertainty about the geographical distribution of fossil-fuel emissions. We have explored the uncertainty of gridded emissions estimates in the United States in order to develop an approach for understanding the uncertainty in the spatial distribution of emissions generally.

Our analyses of the spatial distribution of CO₂ emissions in the United States enforce the point that emissions from large point sources are a major portion of total emissions and that overall uncertainty would be much reduced with more complete and more certain data on the magnitude and location of large point sources. In *Woodard et al.* [2015], we have derived a new statistic, PSUM, that we believe is useful in dealing with the locational uncertainty of large point sources. We treat PSUM as a standard deviation of the uncertainty contributed by the spatial location of large point sources.

For the distribution of emissions from small and areal sources, it is common to use a proxy such as population density to approximate the areal distribution of CO₂ emissions. We derive estimates of both the magnitude and spatial uncertainty of data on population density in the United States by calling on data from the U.S. Census Bureau and doing displacement analyses of gridded data. Data on population density make large contributions to uncertainty when sharp boundaries of population density correspond closely with grid cell boundaries, a situation that occurs infrequently for 1° grid cells but can be expected more commonly as grid cell size is decreased.

A dominant and persistent source of uncertainty is uncertainty in how well the distribution of a proxy represents the distribution of CO₂ emissions. Population density turns out to be a useful proxy in that we have data at multiple scales for the United States (e.g., national and state) and we can use these data to begin to construct an estimate of proxy uncertainty. At 1° resolution, proxy uncertainty is the most important source of uncertainty in most grid cells, those without large point sources. The importance of proxy uncertainty could be substantially reduced by the ability to associate more of total emissions to large point sources. In our analysis, we have used data on emissions from fossil-fuel power plants but not from other large industrial facilities. Having data on progressively smaller sources would also bring a larger fraction of total emissions into the point source category and reduce the importance of proxy error for the non-point source emissions. This observation also emphasizes that the treatment here is constrained by the availability of data on large point sources. For long time series, the availability of data on large point sources is limited or non-existent and proxy uncertainty will dominate. Proxy uncertainty might be constrained by using periodic, labor-intensive, sectoral and bottom-up inventories to essentially help calibrate the proxy relationship. This approach is being pursued by *Rayner et al.* [2010] in merging two proxies (population density and night lights) for comparison against the Vulcan inventory for the United States [*Gurney et al.*, 2009].

All of these discussions emphasize that uncertainty is very scale-dependent. At progressively finer scales, a larger fraction of large point sources will lie near cell boundaries with uncertainty whether the reported cell location is accurate. Small locational errors also become more important for areal sources at fine scale, and we become less certain that distribution of the proxy accurately represents distribution of CO₂ emissions. Because of the locational uncertainty and the possibility that both large and small sources are displaced across cell boundaries, it is clear that the uncertainties of nearby cells are not independent. If a large point source has a possibility that it is not where it was reported, then there is a possibility that it is where it was not reported. This interdependence will be less important as grid size increases and the importance of spatial error decreases. The results of this analysis deal only with the uncertainty of emissions from a specific geographic area (a given grid cell) and do not consider the correlations among grid cells.

We have developed a multi-faceted methodology for characterizing the uncertainty associated with gridded CO₂ emissions data sets. Each component of uncertainty within a data set can be calculated using these methods and combined to produce a final uncertainty mapping for gridded data. The analyses suggest that at 1° latitude/longitude resolution, the current uncertainty (one standard deviation) by grid space in the United States, for our derivative data set, is on the order of $\pm 150\%$. Taking this analysis to a global scale will require additional analysis to characterize spatial uncertainty for each country or group of similar countries. We anticipate that the uncertainty will be larger for many countries where data on large point sources and the distribution of population are less well documented. While data users need to appreciate the uncertainty in these data sets, the best data are probably suitable for many purposes.

References

- Andres, R. J., G. Marland, I. Fung, and E. Matthews (1996), A one degree by one degree distribution of carbon dioxide emissions from fossil-fuel combustion and cement manufacture, 1950–1990, *Global Biogeochem. Cycles*, 10(3), 419–429, doi:10.1029/96GB01523.
- Andres, R. J., et al. (2012), A synthesis of carbon dioxide emissions from fossil-fuel combustion, *Biogeosciences*, 9, 1845–1871, doi:10.5194/bg-9-1845-2012.
- Andres, R.J., T.A. Boden, and G. Marland, 2013. *Annual Fossil-Fuel CO₂ Emissions: Mass of Emissions Gridded by One Degree Latitude by One Degree Longitude*, Carbon Dioxide Inf. Anal. Cent., Oak Ridge Natl. Lab., U.S. Dep. of Energy, Oak Ridge, Tenn. 10.3334/cdiac/ffe.ndp058.2013
- Andres, R. J., T. A. Boden, and D. Higdson (2014), A new evaluation of the uncertainty associated with CDIAC estimates of fossil fuel carbon dioxide emission, *Tellus B*, 66, 23616, doi:10.3402/tellusb.v66.23616).
- CarMA (2013), Carbon monitoring for action. Retrieved from <http://carma.org>
- Gurney, K. R., D. L. Mendoza, Y. Zhou, M. L. Fischer, C. C. Miller, S. Geethakumar, and S. de la Rue du Can (2009), High resolution fossil-fuel combustion CO₂ emissions fluxes for the United States, *Environ. Sci. Technol.*, 43(14), 5535–5541, doi:10.1021/es900806c.
- Hutchins, M. G., J. D. Colby, G. Marland, and E. Marland (2016), A comparison of five high-resolution spatially-explicit fossil fuel carbon dioxide emissions inventories, *Mitig. Adapt. Strat. Global Change*, doi:10.1007/s11027-016-9709-9.
- Nassar, R., L. Napier-Linton, K. R. Gurney, R. J. Andres, T. Oda, F. R. Vogel, and F. Deng (2013), Improving the temporal and spatial distribution of CO₂ emissions from global fossil fuel emissions data sets, *J. Geophys. Res.*, 118, 917–933, doi:10.1029/2012JD018196.
- Oak Ridge National Laboratory (2015), *Landscan*. Oak Ridge Natl. Lab., U.S. Dep. of Energy, Oak Ridge, Tenn. [Available at <http://web.ornl.gov/sci/landscan/>, accessed 5 Aug. 2015]
- Oda, T., and S. Maksyutov (2011), A very high-resolution (1 km × 1 km) global fossil fuel CO₂ emission inventory derived using a point source database and satellite observations of nighttime lights, *Atmos. Chem. Phys.*, 11, 543–556, doi:10.5194/acp-11-543-2011.
- Quick, J., 2014. Carbon dioxide emission tallies for 210 U.S. coal-fired power plants: a comparison of two accounting methods. *J. Air Waste Manage. Assoc.* 64(1): 73-79, 10.1080/10962247.2013.833146.
- Rayner, P.J., M.R. Raupach, M. Paget, P. Peylin, and E. Koffi, 2010. A new global gridded data set of CO₂ emissions from fossil fuel combustion: methodologies and evaluation, *J. Geophys. Res.*, 115, D19306, 10.1029/2009JD013439.
- Schindler, I. (2015), *Measuring smokestack emissions accurately*, NIST Phys. Meas. Lab., Natl. Inst. of Stand. and Technol., 23 Nov. 2015 [Available at <http://www.nist.gov/pml/div685/grp02/smokestack-turbulence.cfm>, accessed 3 Feb. 2016].
- Singer, A. M., M. Branham, M. G. Hutchins, J. Welker, D. L. Woodard, C. A. Badurek, T. Ruseva, E. Marland, and G. Marland (2014), The role of CO₂ emissions from large point sources in emissions totals, responsibility, and policy, *Environ. Sci. Policy*, 44, 190–200, doi:10.1016/j.envsci.2014.08.001.
- U.S. Census Bureau (2011), *Population Distribution and Change: 2000 to 2010: Census Briefs*. U.S. Census Bureau, Suitland, Md. [Available at <http://www.census.gov/prod/cen2010/briefs/c2010br-01.pdf>.]
- U.S. Department of Energy/Energy Information Administration (2013), *Energy-Related Carbon Dioxide Emissions at the State Level*. U.S. DOE/EIA, Washington, D.C. [Available at: <http://www.eia.gov/environment/emissions/state/analysis/>]
- U.S. Environmental Protection Agency (2014), *Clean Energy: eGRID, Ninth Edition With 2010 Data*. U.S. Environ. Prot. Agency, Washington, D.C. [Available at: <http://www.epa.gov/cleanenergy/energy-resources/egrid/>.]
- U.S. Environmental Protection Agency (2015a), *Greenhouse Gas Reporting Program (GHGRP) 2011*. U.S. Environ. Prot. Agency, Washington, D.C. [Available at: <http://www2.epa.gov/ghgreporting>.]
- U.S. Environmental Protection Agency (2015b), *Inventory of U.S. Greenhouse Gas Emissions and Sinks (1990–2013)*. U.S. Environ. Prot. Agency EPA 430-R-15-004, Washington, D.C. [Available at: <http://www3.epa.gov/climatechange/Downloads/ghgemissions/US-GHG-Inventory-2015-Main-Text.pdf>.]
- Woodard, D., M. Branham, G. Buckingham, S. Hogue, M. Hutchins, R. Gosky, G. Marland, and E. Marland (2015), A spatial uncertainty metric for anthropogenic CO₂ emissions, *Greenhouse Gas Meas. Manage.*, doi:10.1080/20430779.2014.1000793.